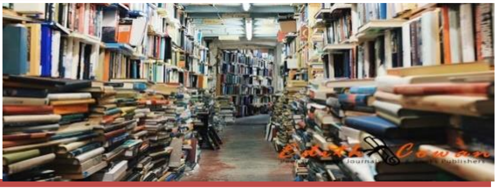
Peer Reviewed Journals & Books Publishers

Edith Cowan Journal of Education
ISSN: 2790-055X
Vol.2, Issue No.1, pp 52-60.



Statistical significance and effect size in education research: two sides of a coin

[1]Trevor Haddish
[1]Faculty of Education and Applied Science, University of the Philippines Diliman,
Corresponding author's e-mail: haddishtrevor@gmail.com

ARTICLE INFO

Article history:
Received Date: 10th Nov 2021
Revised Date: 11th April 2022
Accepted Date: 18th Dec 2022

Keywords:
Effect-size measures, research practice, statistical significance testing

ABSTRACT

T In education research, statistical significance and effect size are 2 sides of 1 coin; they complement each other but they do not substitute for each other. Good research practice requires that, to make sound research decisions, both sides should be considered. In a simulation study, the sampling variability of 2 popular effect-size measures (d and R2) was examined. The variability showed that what is statistically significant may not be practically meaningful, and what appears to be practically meaningful could have been the result of sampling error, thus not trustworthy. Some practical guidelines are suggested for combining the 2 sources of information in research practice.

Background

In education research, statistical significance testing has received many valid criticisms in recent years primarily because the outcome of statistical significance testing relies too heavily on sample size, and the issue of practical significance is often ignored. Consequently, too much reliance on statistical significance testing often limits understanding and applicability of research findings in education practice. Effect size has been proposed as a supplement or an alternative to statistical significance testing; it has become increasingly popular. Some education researchers, however, may not be aware that, by itself, effect size can also be misleading because sample size influences the sampling variability of an effect-size measure. Through a Monte Car10 experiment, I show that statistical significance testing and effect size are two related sides that together make a coin; they complement each other but do not substitute for one another. Good research practice requires that, for making sound quantitative decisions in education research, both sides should be considered. To lay a foundation for the discussion in this article, I first reviewed some major issues related to statistical significance testing and effect-size measures. e and effect size are 2 sides of 1 coin; they complement each other but they do not substitute for each other. Good research practice requires that, to make sound research decisions, both sides should be considered. In a simulation study, the sampling variability of 2 popular effect-size measures (d and R2) was examined. The variability showed that what is statistically sigd5cant may not be practically meaningful, and what appears to be practically meaningful could have been the result of sampling error, thus not trustworthy. Some practical guidelines are suggested for combining the 2 sources of information in research practice. Key words: effect-size measures, research practice, statistical significance testing n education research, statistical significance testing has I received many valid criticisms in recent years primarily because the outcome of statistical significance testing relies too heavily on sample size, and the issue of practical significance is often ignored. Consequently, too much reliance on statistical significance testing often limits understanding and applicability of research findings in education practice. Effect size has been proposed as a supplement or an alternative to statistical significance testing; it has become increasingly popular. Some education researchers, however, may not be aware that, by itself, effect size can also be misleading because sample size influences the sampling variability of an effect-size measure. Through a Monte Car10 experiment, I show that statistical significance testing and effect size are two related sides that together make a coin; they complement each other but do not substitute for one another. Good research practice requires that, for making sound quantitative decisions in education research, both sides should be considered. To lay a foundation for the discussion in this article, I first reviewed some major issues related to statistical significance testing and effect-size measures.

Literature Review

Statistical Significance Testing

Use of statistical significance testing in research. There have been different misconceptions about what significance testing is and what it is not (Shaver, 1993). For this article, one should have a good understanding about the basic purpose of statistical significance testing in quantitative research and about what information statistical significance testing provides for education researchers.

The fundamental concept underlying statistical significance testing is sampling variation. From a population with a known parameter (e.g., known population mean), sample statistics (e.g., observed means of multiple samples) will vary around the population parameter to a certain extent. Because of the sampling variability, the difference between an observed sample statistic (e.g., sample mean) and the population parameter (i.e.. population mean) does not necessarily indicate that the sample does not belong to the population. For example, if the mean of a random sample (N = 20) is observed to be 68, could this sample statistic have occurred because of sampling variability (i.e., by chance) if the population mean is 80? A statistical significance test can be conducted to evaluate the viability of the hypothesis that the sample with a mean of 68 could have been drawn from a population with a mean of 80. That evaluation is done by assessing how likely the difference between the observed sample statistic and the known population parameter could have occurred as the result of chance, that is, random sampling variation. In other words, statistical significance testing answers the question: What is the probability of obtaining an observed sample statistic (e.g., mean of 68) when the population has a known parameter value (e.g., population mean of 80)?

Assume that two treatment conditions (A and B) exist; (e.g., A represents a new instructional approach for teaching mathematics and B represents the conventional instructional approach currently in use). The researcher is interested in knowing if A is more effective than B in teaching mathematics (RH: research hypothesis, A is better than B). To help decide if the RH can be supported, the researcher set up another hypothesis (NH: null hypothesis of no difference) that A and B are equally effective, that is, students under A and B will learn equally well.

Because of sampling variation, even if A is the same as B in terms of effectiveness, the sample under A may have higher sample mean than the sample under B. So the observation that students under A performed better than those under B does not necessarily mean that Method A is more effective than B because sampling error has not been ruled out as one possible explanation for the observed difference between the two samples.

If A and B treatments are the same (NH: of no difference), a small performance difference between A and B samples is more likely to occur by chance than is a large performance difference. When the difference between the two samples becomes sufficiently large relative to the random sampling variation, however, one begins to doubt that A and B are equally effective. In that case, it should be highly unlikely to observe the large performance difference between the two samples. The question becomes: How much higher should the mean of the Method A sample be than that of the Method B sample before one can determine with reasonable confidence that the observed difference is not due to sampling variability (i.e., chance)? Once one decides statistically that the sampling variability is no longer a viable explanation for the observed difference, the NH will be rejected in favor of the RH (Method A is more effective than Method B). The rejection of NH constitutes evidence for supporting the RH because a statistical significance test helps to eliminate sampling error, or chance, as a viable explanation for the observed difference between the two samples.

In the statistical significance testing, I assessed the probability of obtaining the sample data (0) if the null hypothesis (HJ is true, that is, p(D I H,). If p(D I H,) is sufficiently small (e.g., smaller than .05 or .Ol), the null hypothesis will be considered not viable and will be rejected. The rejection of the null hypothesis indicates that the random sampling variability is the unlikely explanation for the observed statistical results, but it does not generally show the importance of obtained statistical results. Regarding the example of A and B methods in teaching mathematics, rejection of the null hypothesis (A and B are equally effective in teaching mathematics) simply means that, given the observed magnitude of difference between the two samples, it is highly unlikely that sampling error could have been the cause for the observed difference. As a result, one concludes that A and B are probably not equally effective. That conclusion, however, does not provide a clear indication about how much more effective Method A is than Method B in the practical sense. Unfortunately, the meaning of statistical significance testing as discussed here has sometimes been lost, and the importance of statistical significance tends to be grossly exaggerated in education research practice.

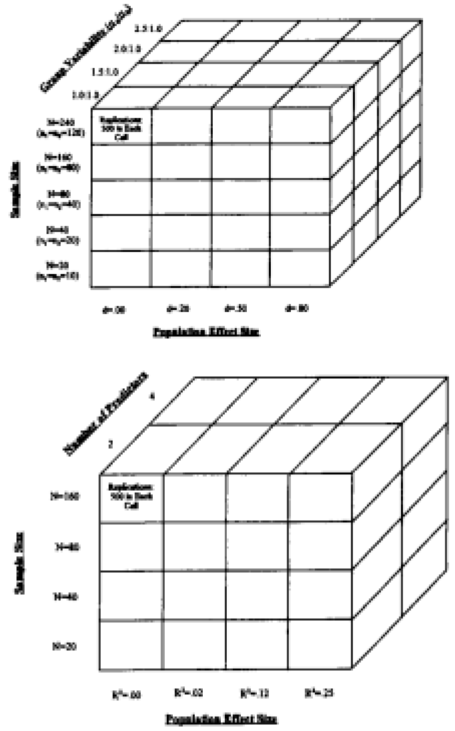
Materials and Methods

Although theoretical sampling distributions of some popular effect-size measures have been known (e.g., see Hedges & Olkin, 1985 ford, Glass & Hopkins, 19% for R2), I used an empirical approach in this article to provide more intuitive discussion about the relevant issues. I conducted a Monte Carlo experiment to simulate different data conditions under which both effect-size measures and statistical significance testing outcomes were obtained and later analyzed.

Design

In this article, I used the two most widely known effect size measures: d (standardized mean difference) and R2 (proportion of variance accounted for). The two effect-size measures are generally known to researchers who have been exposed to the concept of effect size. The literature review of several psychology journals by Kirk (1 996) indicates that R2 is by far the most frequently reported effect-size measure, probably because it is routinely reported in regression or general linear-model procedures. The meta-analysis work by Glass (1976) undoubtedly contributed to the popularity of d as the effect-size measure. I generated samples from two statistical populations with known population parameters to evaluate stan- two group mean difference (4; see Figure 1. I considered three factors in the Monte Carlo simulation design: (a) four levels of population effect size (d = .OO, .20, SO, and .80, respectively) that correspond to zero, small, medium, and large effects as suggested by Cohen (1988, chapter 2); (b) five levels of sample-size conditions (N = 20, 40, 80, 160, 240); and (c) four conditions of group variability ratio (represented by the population standard deviations) between two populations (ol/ o2 = 1, 1.5, 2, and 2.5, respectively). For the fully crossed design, the three factors yielded 80 (4 x 5 x 4) cells. I conducted 500 replications within each cell; the total number of replications in this Monte Car10 experiment for evaluating d were 40,000 (500 x 80).

Figure 1. Study Design for Effect-Size Measures of *d* and *R*²



I used regression models to evaluate R2 (proportion of variance accounted for). I considered three factors in the design: (a) four levels of population effect size (R2 = .00, .02, .12, and .25, respectively), which approximately correspond to zero, small, medium, and large effects as suggested by Cohen (1988, chapter 9); (b) four levels of sample size conditions (N = 20,40, 80, and 160, respectively); and (c) two conditions for the number of predictors (k = 2 and 4, respectively), with the correlation among the predictors set at r = .lo. The fully crossed design of the three factors called for 32 cells (4 x 4 x 2). With 500 replications within each cell, the total number of replications for the experiment was 16,000 (32 x 500). The designs for evaluating d and R2 are presented graphically in Figure 1.

Data

I attained data generation by using the SAS normal data generator. Multivariate normal data for regression models were simulated with the matrix decomposition procedure (Kaiser & Dickman, 1962). I accomplished all sample data generation, sample effect-size calculation, and statistical significance testing through the Interactive Matrix Language (PROC IML) of the SAS system (SAS Window Version 7.0). I did not consider data non-normality in this study. As a result, the influence of data non-normality on both effect-size measures and statistical significance test outcomes was not assessed.

Results and Discussion

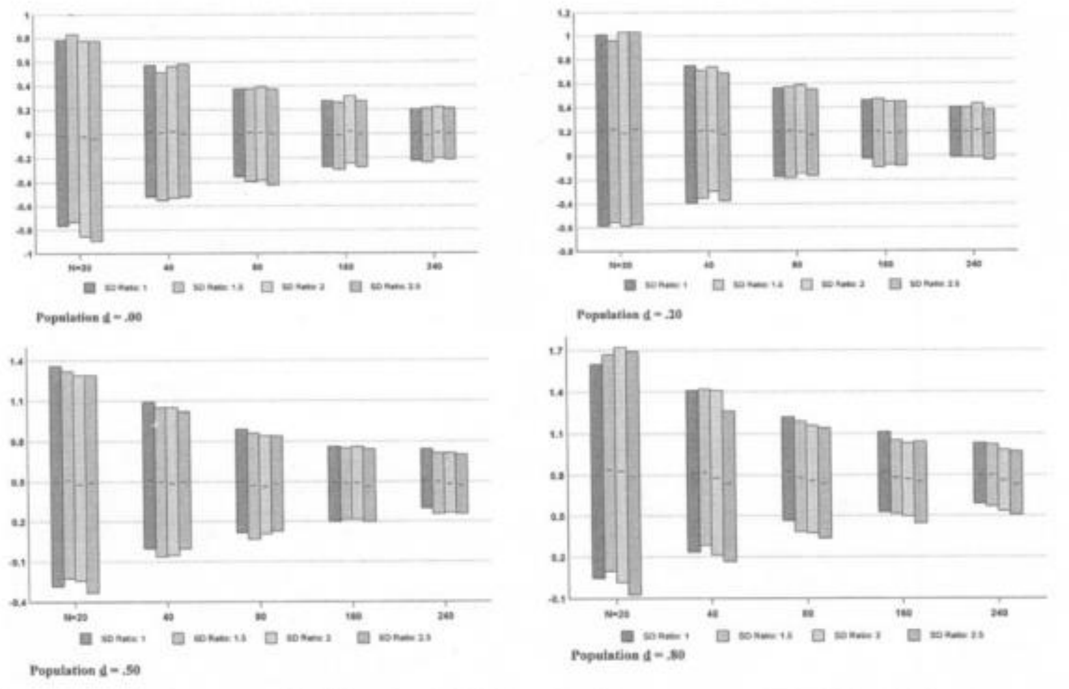
Figure 2 graphically describes the sampling variability of the effect-size measure of d for four conditions of population effects: zero, small, medium, and large (population d = .00, .20, SO, and .80, respectively). In addition to sample size conditions, the four conditions of group variability ratio(o,/ 02) are also presented (ol/ o2 =1, 1.5, 2, and 2.5). In Figure 2, a high-low bar represents the 90% confidence interval of sample d for a condition of sample size and for a group variability ratio (ratio of the standard deviations of two groups), and a short horizontal line within a bar represents the mean of 500 sample ds. Several observations can be made from Figure 2. First, sample effect-size measured appears to be an unbiased1 estimate of population d. The lack of bias of sample d is obvious because over repeated sampling, the mean of sample d is very close to the known population value specified in the Monte Car10 experiment (population d = .OO, .20, SO, and 30, respectively) under most data conditions. However, a larger discrepancy between the two population standard deviations (SD ratio) causes some minor degree of downward bias of sample d, and this is especially obvious under the condition of population d = 30.

Second, there is considerable sampling variability of sample effect size d. For example, under the condition of population d = .OO (i.e., two samples drawn from the same population, thus no real difference between the two samples), for small-size condition such as N = 20 (nl = n2 = lo), the 90% confidence interval almost covers the range from -30 to +.80. In other words, for that sample-size condition, when two samples were drawn from the same population, and consequently, there was no real difference between the two groups, I could have obtained a large effect size (i.80) just by chance (i.e., sampling error). Even when sample size was increased to N = 80 (n, = n2 = 40), probably a moderate sample size for many experimental designs, I still could have obtained sample effect size almost as large as k.40 (moderate effect) by chance.

Third, the extent of sampling variability is obviously influenced by sample size. With the increase of sample size, the sampling variability of sample d, as represented by the 90% confidence intervals, shows a clear trend of becoming gradually smaller under all the conditions of population effect size (zero, small, medium, and large). That trend indicates that, if there are two identical effect sizes (e.g., moderate effect of d = .40) from two different studies involving different sample sizes (e.g., one is based on sample size of 40 [nl = n2 = 201, and the other is based on N = 160 [n, = n2 = 80]), the one based on the larger sample size is more trustworthy because such an effect size is very unlikely to have occurred because of sampling error. That result indicates that the use of effect size measure should take sample size into consideration.

The sampling variability of another major type of effect size measure is shown in Figure 3-the measure of association strength as represented by R2. Because sample R2 is widely known to have upward bias, I also reported one form of bias-corrected R2 (adjusted R2 obtainable from SAS or SPSS regression procedure) in Figure 3. The sampling variability of the R2 and adjusted R2 is represented by the 90% confidence interval bar; the mean R2 based on 500 replications is represented by the short horizontal line within each confidence interval bar.

Figure 2. Confidence Intervals (90%) of Sample Effect Size *d* for Four Conditions of Population Effects



Population d = .00

Population d = .20

Population d = .50

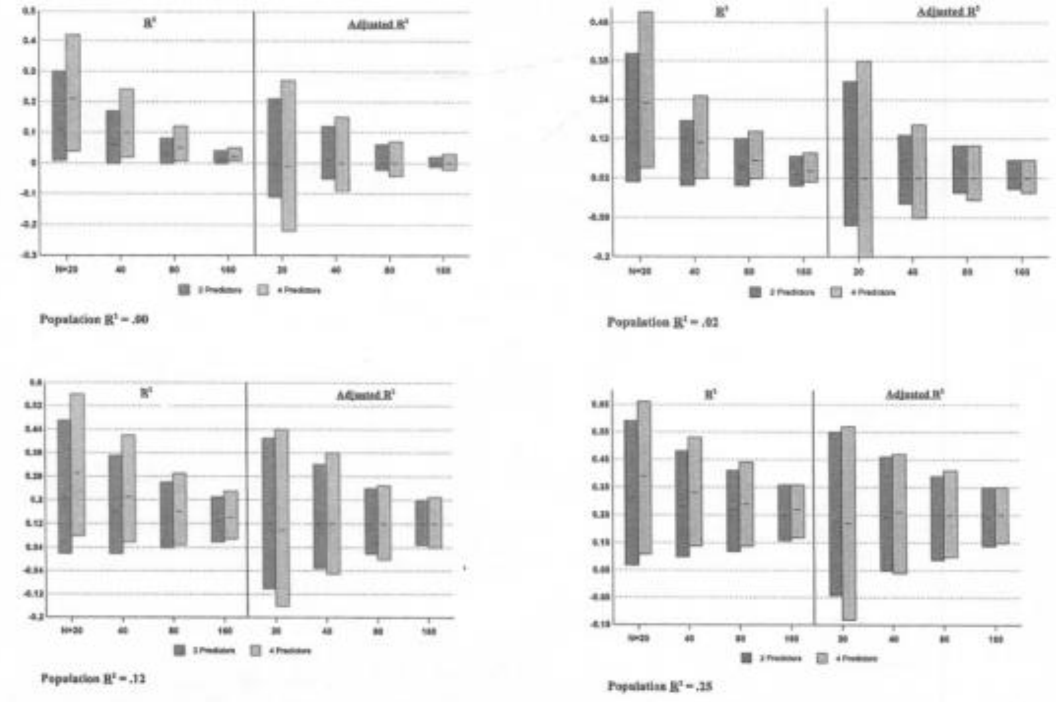
Population d = .80

In addition to some common observations already discussed for the effect-size measured in Figure 2, several observations unique for sample R2 in Figure 3 can be made. First, whereas measured in Figure 2 has been shown to be an unbiased estimator of population d, sample R2 has obvious upward bias, as indicated by the position of mean R2 (short horizontal line within each 90% confidence interval) that is consistently, and sometimes considerably, above the population R2 under all conditions. Bias correction, however, has worked well because the means of all sample adjusted R2s are very close to the population R2 value.

Second, sample R2 from the four-predictor regression model has more upward bias than that from the two-predictor regression model. That finding is expected because under the same sample-size condition, the ratio of sample size to the number of predictors (Nlp) is smaller for the four-predictor model than that for the two-predictor model. As is widely known in regression analysis, it is the ratio, rather than sample size per se, that largely determines the stability of regression analysis outcomes (Stekens, 1996; Yin & Fan, in press).

Both sample R2 and adjusted R2 show considerable sampling variability, which decreases as the sample size increases. The considerable sampling variability may make obtaining a medium and even large effect-size measure by chance relatively easy, even when the population effect size is zero or very small (R2 = .02). For example, for population R2 = .02 (very small effect) and for the four-predictor regression model, the upper 90% confidence limit of sample R2 reaches as high as .46 (very large effect) for N = 20 and about .25 (large effect) for N = 40. That degree of sampling variability influenced by sample size (or Nlp ratio) highlights the need for effect size to be considered in combination with sample size; used by itself, sample effect-size measure could be misleading.

Figure 3. Confidence Intervals (90%) of Sample Effect Size R^2 for Four Conditions of Population Effects



Population R² = .00

Population R² = .02

Population R² = .12

Population R² = .25

Table 1 reports the percentages of statistically significant tests under different population effect-size and sample-size conditions. When the population effect is zero, approximately 5% of tests are statistically significant (underlined entries in the table), close to the specified nominal Type I error rate (a level). When population effect size is not zero, Table 1 entries represent the power of the statistical tests in rejecting the false null hypothesis. The tests can adequately detect the population effect (adequate statistical power is defined to be about .80 [Stevens, 19961) only when the population effect is moderate to large (d = .50, 30) and the sample size is not small (N2 40; see boldface entries in Table 1).

One does not want to trust something that could have occurred by chance (Type I error); Table 1 shows that, when there is true effect, statistical tests may cause concern of Type I1 error. In other words, one may conclude that there is no effect when, in fact, there is. Balancing the two opposite logical errors requires the researcher (a) to understand the consequences of Type I and I1 errors, respectively; (b) to consider effect-size measure; and (c) to make decisions accordingly. Practical guidelines for combining statistical significance testing and effect-size measure in research practice are offered in Table 2. The content of Table 2 is self-explanatory; therefore, no explanation or discussion is required here.

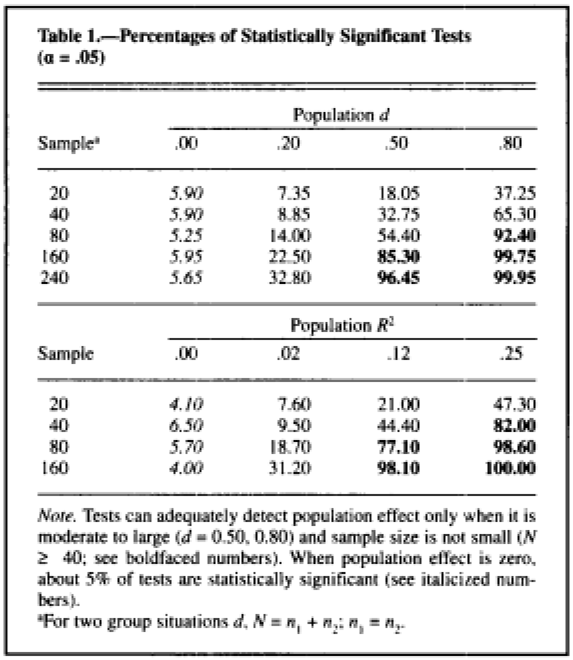


Table 1.—Percentages of Statistically Significant Tests (α = .05)

	Population *d*			
Sample[a]	.00	.20	.50	.80
20	*5.90*	7.35	18.05	37.25
40	*5.90*	8.85	32.75	65.30
80	*5.25*	14.00	54.40	**92.40**
160	*5.95*	22.50	**85.30**	**99.75**
240	*5.65*	32.80	**96.45**	**99.95**

	Population R^2			
Sample	.00	.02	.12	.25
20	*4.10*	7.60	21.00	47.30
40	*6.50*	9.50	44.40	**82.00**
80	*5.70*	18.70	**77.10**	**98.60**
160	*4.00*	31.20	**98.10**	**100.00**

Note. Tests can adequately detect population effect only when it is moderate to large (d = 0.50, 0.80) and sample size is not small ($N \geq$ 40; see boldfaced numbers). When population effect is zero, about 5% of tests are statistically significant (see italicized numbers).

[a] For two group situations d, $N = n_1 + n_2$; $n_1 = n_2$.

Conclusion

In this article, I attempted to show that statistical significance testing and effect size are two related sides of the same coin; they complement each other but they do not substitute for each other. Good research practice requires that both sides should be taken into account to reach sound quantitative research decisions. The Monte Car10 experiment showed empirically that there is considerable variability of sample effect-size measure and that the extent of such variability is influenced by sample size. Because of the sampling variability of an effect-size measure, what appears to be practically meaningful effect size may be the result of sampling error, and, consequently, is not trust-worthy. Statistical significance testing and effect-size measure serve different purposes; the sole reliance on either may be misleading. Some practical guidelines (see Table 2) are suggested for combining statistical significance testing and effect-size measure to make decisions in research practice

Reference List

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Carver, R. (1978). The case against statistical significance testing. Harvard Education Review, 48, 378-399. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd. ed.). New York:

Erlbaum. Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & Psy

Fowler, R. J. (1985). Point estimate and confidence intervals in measures of association. Psychological Bulletin, 98. 160-165.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Education Researcher; 5, 3-8.

Glass, G. V., & Hopkins, K. D. (1996). Statistical methods in education andpsychology (3rd ed.). Boston, MA: Allen & Bacon.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. Psychometrika, 27, 179-182.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Education and Psychological Measurement. 56, 746-759.

Levin, J. R. (1993). Statistical significance testing from three perspectives. The Joumal of Experimental Education, 61, 378-382.

Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and chology, 5, 75-98. analyzing data: A model comparison perspective. Belmont, CA: Wadsworth.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806834.

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance. The Journal of Experimental Education. 61, 383-387.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. The Journal of Experimental Education, 61, 293-316.

Snyder, P., & Lawson, S. (1993). Effect size estimates. The Journal of Experimental Education, 61, 334-349.

Stevens, J. (1996). Applied mulfivariate statistics for the social sciences. Mahwah, NJ: Erlbaum.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1993). The use of statistical significance in research: Bootstrap and other alternatives. The Journal of Experimental Education, 61, 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Education Researcher; 25, 26-30.

Yin, P., & Fan, X. (in press). Estimating RZ shrinkage in multiple regression: A comparison of different analytical methods. The Journal of Experimental Education.

Wilkinson, L., &The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.